IEICE TRANS. INF. & SYST., VOL.E94–D, NO.10 OCTOBER 2011

PAPER Special Section on Information-Based Induction Sciences and Machine Learning

Augmenting Training Samples with a Large Number of Rough Segmentation Datasets

Mitsuru AMBAI[†a)], ***Member and*** **Yuichi YOSHIDA**[†], ***Nonmember***

SUMMARY We revisit the problem with generic object recognition from the point of view of human-computer interaction. While many existing algorithms for generic object recognition first try to detect target objects before features are extracted and classified in processing, our work is motivated by the belief that solving the task of detection by computer is not always necessary in many practical situations, such as those involving mobile recognition systems with touch displays and cameras. It is natural for these systems to ask users to input the segmentation data for targets through their touch displays. Speaking from the perspective of usability, such systems should involve *rough* segmentation to reduce the user workload. In this situation, different people would provide different segmentation data. Here, an interesting question arises – if multiple training samples are generated from a single image by using various segmentation data created by different people, what would happen to the accuracy of classification? We created "20 wild bird datasets" that had a large number of rough segmentation datasets made by 383 people in an attempt to answer this question. Our experiments revealed two interesting facts: (i) generating multiple training samples from a single image had positive effects on classification accuracies, especially when image features including spatial information were used and (ii) augmenting training samples with artificial segmentation data synthesized with a morphing technique also had slightly positive effects on classification accuracies.

key words: *Interactive recognition, generic object recognition*

1. Introduction

Algorithms for generic object recognition generally first detect target objects before features are extracted and classified in processing. Since an input image includes various visual information, the target detection task plays an important role in removing insignificant image signals such as background components. A sliding window is a classical way of detecting targets. This approach tests predefined bounding boxes on the targets using an object classifier. Although exhaustive searches for possible bounding boxes are apt to increase computational costs, it has been demonstrated that a cascaded classifier [1], [2] and a branch and bound search approach [3] can effectively reduce computational costs. Another approach makes use of the recent advantages of the segmentation technique [4], [5] to identify the region containing the target object. Gu et al. [6] focused on the subdivided regions of an input image as perceptually meaningful entities for the purpose of recognizing objects. They divided an input image into small parts and used a "bag of regions" representation to recognize the objects. The region-based approach naturally incorporates shape and scale information into a feature vector without being affected by clutter from outside the regions. This is a great advantage in generic object recognition.

However, fully automatic object segmentation by computer is still a complex problem because of ambiguity in the object boundary. Let us consider some body parts of an avian animal such as its head, legs, abdomen, chest, and wings. No one can determine the "exact" boundaries of these parts, because their boundaries are not precisely defined. This means there is no ground truth for segmentation. Despite such difficulties, most methods of segmentation determine the boundary of regions based on edge intensities without taking into account their conceptual meanings. The segmentation of natural objects is a still challenging problem that remains unsolved in machine vision.

2. Motivation and Contribution

Our work was motivated by a belief that *solving segmentation tasks by computer is not always necessary in many practical cases*. Let us take an example of a recognition system for wild birds using a mobile device equipped with a touch display and a camera. It is natural in this situation to delegate the task of segmentation to users. Human visual perception is still superior to that of computers. Even if users do not know the name of the target, they can easily draw a contour of the target and even boundaries of their parts with a touch display. The computer can then skip image segmentation and only concentrate on feature extraction and the classification. This means that users and the computer can share the recognition task. Users can help detect the target, and the computer can provide the name of the target.

From the perspective of usability, the system should not require users to draw closed contours in detail. Moreover, as was pointed out earlier, it is impossible to define precise boundaries in the first place because of the problem with ambiguity. For these reasons, the recognition system should ask the user to input rough segmentation data. Figure 1 shows examples of roughly segmented regions drawn by 10 different people. We asked them to draw four closed curves that included four body parts of a bird: its entire body, head, wing and torso (i.e., abdomen and chest). As can be seen from the figure, they enclosed the regions in various shapes depending on their own interpretations. All different segmentation data provide different feature vectors

Manuscript received December 28, 2010.
Manuscript revised May 4, 2011.

[†]The authors are with Denso IT Laboratory, Inc., Tokyo, 150–0002 Japan.
a) E-mail: manbai@d-itlab.co.jp

DOI: 10.1587/transinf.E94.D.1880

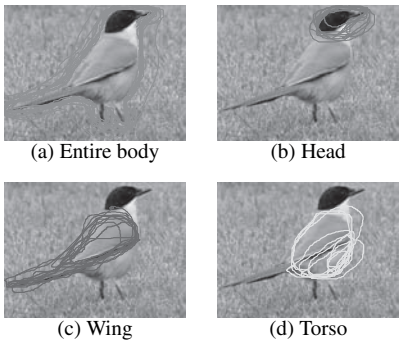

(a) Entire body (b) Head
(c) Wing (d) Torso

Fig. 1 Ten rough segmentations drawn by different people who segmented data differently for single image.

even from the same image. This characterizes the setting for our problem.

Here, we propose a hypothesis that if training samples are expanded by adopting diverse segmentation data, higher classification accuracies will be achieved. While previous approaches such as those by [7], [8] have generated only a single training sample from a single image, our approach generates multiple training samples from a single image using different segmentation results made by different people. Our approach represents the ambiguity of segmentation as diverse training samples. The problem with ambiguity is expected to be solved in the learning process.

To test this hypothesis, we prepared "20 wild bird datasets" with rough segmentation data that were composed of four types of regions: the entire body, head, wing, and torso of a bird. Ten different people independently prepared the segmentation data for each single image. Seven kinds of image features were tested through our experiments that proved two interesting facts:

1. Generating multiple training samples from a single image based on diverse segmentation datasets had positive effects on classification accuracies. The magnitudes of improvements depended on the types of image features. Image features including spatial information such as PHOG [9], [10] and PHOW [10] tended to have greater apparent positive effects than "bags of features" representations.
2. By using synthesized segmentation datasets together with manually made segmentation datasets, the classification accuracies were slightly improved. Synthesized segmentation datasets could be obtained by using a morphing technique that blended the two different segmentation data made by people.

3. Related Works

3.1 Existing Methods

Classification accuracy strongly depends on the number of training samples. A larger number of training samples generally results in higher classification accuracy. Since collecting a sufficient number of training samples is problematic, learning an object model from a small number of training images has attracted a great deal of attention.

Representative works with the above motivation have been proposed in [11], [12]. Fei-Fei et al. [11] created visual models of object categories from just a few images, i.e., less than or equal to five images per single category. They used the generic knowledge derived from visual models that had already been learned in advance. Lampert et al. [12] tackled a more challenging problem where no training samples of a target class were available. They proposed attribute-based classification to solve this problem. Their method was based on the idea that certain kinds of attributes are shared even with different categories. Image feature information is transferred via the attributes from categories with training samples to categories without training samples.

While we shared an awareness of the issues with [11], [12], our approach utilized a technique that is often used in the field of image pattern recognition [13], [14]. Ranzato et al. [13] augmented the training image samples with new image samples obtained by elastically distorting the original training samples and they achieved the best classification rate for the MNIST dataset [15], which is a widely used dataset of handwritten digits. Lauer et al. [14] used affine transformation to deform the original training samples and obtained comparable results to [13]. This work revealed that synthesizing new datasets from the original datasets effectively improved classification accuracy. Our approach also belongs to these kinds of studies.

From the perspective of human-computer interaction, the system proposed by Zou et al. [16] is similar to our work. Their system asks users to input the roughly shaped model of a target to recognize flowers. Although their concept is similar to ours in terms of sharing the recognition task between users and a computer, many more improvements can be made to the part to extract features. They used simple image features such as color histograms that did not have sufficient discriminative power to enable generic object recognition. Researchers in the field of human-computer interaction and multimedia generally tend to avoid discussion on image features that are vitally important for generic object recognition.

3.2 Difference between Our Method and Existing Methods

We attempt to augment training samples as well as [13] and [14]. It can be safely said that the main difference between our method and the existing methods is a way how to deform a feature vector of a target. The existing methods, so-called elastic distortion [13] and affine distortion [14], geometrically deform pixel patterns. On the other hand, our method deforms a boundary of the target in accordance with various segmentation datasets provided by different people.

The reason why we preferred the boundary deformation to the elastic and the affine deformation is that we intend to incorporate human-factor into augmenting training samples. For example, as can be seen from Fig. 1, people

differently recognized boundaries of the torso and provided enormously varied boundaries. On the other hand, all of the people gave relatively similar boundaries to the head. This is caused by ambiguity of object boundaries. The level of the variation of boundaries depends on parts of targets and individuals. It is difficult to cope with the ambiguity issue of boundaries by the simple systematic approaches such as the elastic and the affine distortion, because they don't take into account any of the human-factor. It is needless to say that if the pixel displacements by the elastic and affine deformation are too large to preserve original signals, the classification accuracy will decrease rather than increase. Therefore it is very important to choose good deformation parameters which provides naturally deformed image sets. However, in our problem setting, it is very difficult to find such good parameters because the deformation magnitude of boundaries greatly depends on parts of targets and individuals. By collecting real segmentation data from people in advance, we can naturally model the variation of boundaries. This leads to good classification accuracy. In summary, our approach revisits the conventional approach of increasing training samples from the point of view of human-computer interaction, which leads to solving the ambiguity issue.

In statistics, bootstrapping [17] is often used to estimate parameters by random sampling with replacement from a small number of observations. The concept of the bootstrapping is also similar to that of our method, but does not take the human-factor into consideration.

After this, the rough region boundary drawn by a person will be denoted as a "stroke" to distinguish it from a precise boundary.

4. Datasets and Methodology

4.1 20 Wild Bird Datasets

Some previous image datasets such as the VOC2009 [18] and the Caltech101 [19] have detailed segmentation data that were prepared by people. In contrast, a rough yet a large number of segmentation datasets are necessary to evaluate our idea of interactive recognition.

We created *20 wild bird datasets* with strokes that were created by 383 people. Sample images in our datasets are presented in Fig. 2 and sample strokes are given in Fig. 3. The datasets consist of 20 kinds of wild birds. Each category is composed of 36-112 images, and there are a total of 1,230 images. Ten different individuals made four types of strokes for a single image, i.e., the entire body, head, wing, and torso in turn. Each stroke was a single closed curve and was represented as a two-dimensional point sequence. The average size of the images was approximately 400×490 pixels. The widths ranged from 180 to 640 pixels, and the heights ranged from 116 to 640 pixels. A bird was captured in various poses and on various scales in each of the images and it was not aligned.

These images were downloaded from Google image search and they were manually cropped so as to capture just a single bird in an image. All strokes were collected using the Amazon Mechanical Turk [20] Web service that enabled us to ask a large number of workers to do simple yet intelligent tasks. We sent requests for them to draw the four types of rough strokes on each image. The strokes we obtained in our datasets were made by 383 different workers in total.

4.2 Synthesizing Strokes

A new stroke can be synthesized from two manually made strokes by using a morphing technique. This method, which is similar to elastic distortion [13] and affine transformation [14], makes it possible to increase the number of training samples. The morphing process consists of two steps of (i) obtaining a correspondence between two strokes and (ii) interpolating them.

Scott et al. [21] proposed a method of obtaining correspondences between two closed curves subject to preserving order in the sequence of points. The costs in their formulation of all possible pairing points were expressed as an $M \times N$ cost matrix, where M and N are the numbers of points on the two curves. They solved the assignment problem for the cost matrix by using an efficient dynamic programming algorithm. Here, we denote the numbers of points of two strokes as N_1 and N_2. The size of the cost matrix is $N_1 \times N_2$ and its elements are defined as squared distances among all possible pairing points. Figure 4 (a) shows an example of a matching result.

New points are uniformly inserted onto line segments between the matching points to obtain one-to-one correspondences for every point as seen in Figs. 4 (b) and (c). This generates two resampled strokes that are composed of the same number of points. A new morphed stroke can be generated by linearly interpolating them by using a blending factor in a range from 0 to 1.

4.3 Image Features

This section describes methods of extracting a single feature vector from a single stroke. Here, we introduce seven types of image features without color information for comparison. Three of them are "bags of features" and the rest of them are image features with spatial information.

4.3.1 Bags of Features (without Spatial Information)

A histogram representation of visual words was extracted from an image only inside the stroke according to [22] with slight modifications. Three hundred kinds of visual words were defined based on training images. This provided a 300 dimensional vector for each stroke.

Three types of image features of original SIFT, dense SIFT, and affine SIFT were tested. While these three methods use the same descriptor proposed by [23], their detection processes for points of interest differ. The original SIFT [23] detects feature points with a surrounding circle at the max-

Fig. 2 Sample images of 20 wild bird datasets.

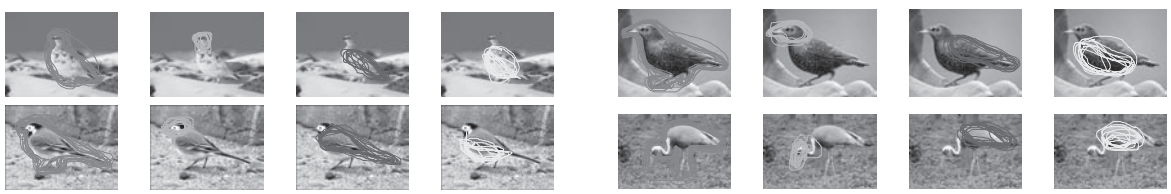
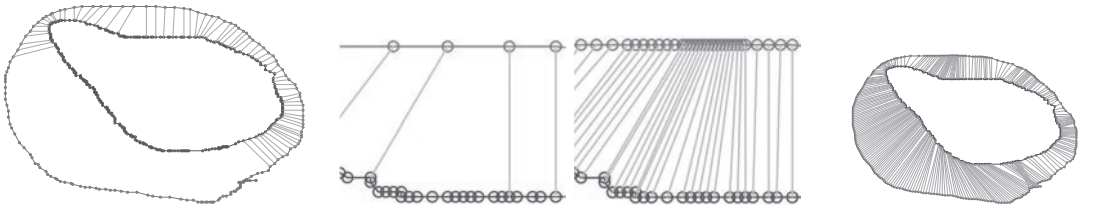
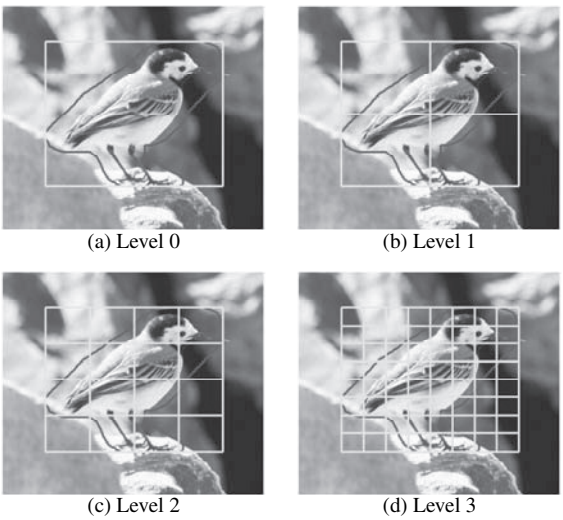


Fig. 3 Sample strokes of entire bodies, heads, wings, and torsos are shown in these figures. Each image has 10 different strokes for each body part. The 10 different strokes have been overlaid on these images.



(a) Matching (b) Resampling (c) One-to-one correspondence

Fig. 4 Morphing process.



(a) Level 0 (b) Level 1

(c) Level 2 (d) Level 3

Fig. 5 Multilevel pyramids for PHOG and PHOW. Bounding box is divided into many cells at each level.

ima/minima of the Difference of Gaussians (DoG) of the given image. Dense SIFT [7] divides the image into blocks of 8×8 pixels, and then every 2×2 blocks are defined as image patches. Descriptors were computed from all the patches. Affine SIFT [24] finds scale and affine invariant points by using a multi-scale representation for the Harris interest point detector. Affine invariant regions around the points are represented as ellipses.

4.3.2 Features with Spatial Information

Pyramid histograms of orientation gradients (PHOG [9]) and pyramid histograms of visual words (PHOW [10]) were also used in our experiments.

PHOG divides a target region into $K \times K$ grids at each level l, where $K = 2^l$ as shown in Fig. 5. The target region is determined as a minimum rectangle without rotation that encloses a stroke. Histograms of the orientation of gradients are calculated from each cell excepting pixels outside the stroke. The gradient angles in the 0 to 360 range are quantized into eight bins.All the histograms of cells that appear in 0 to L levels are concatenated into a single vector with dimensionality $8 \sum_{l=0}^{L} 4^l$.

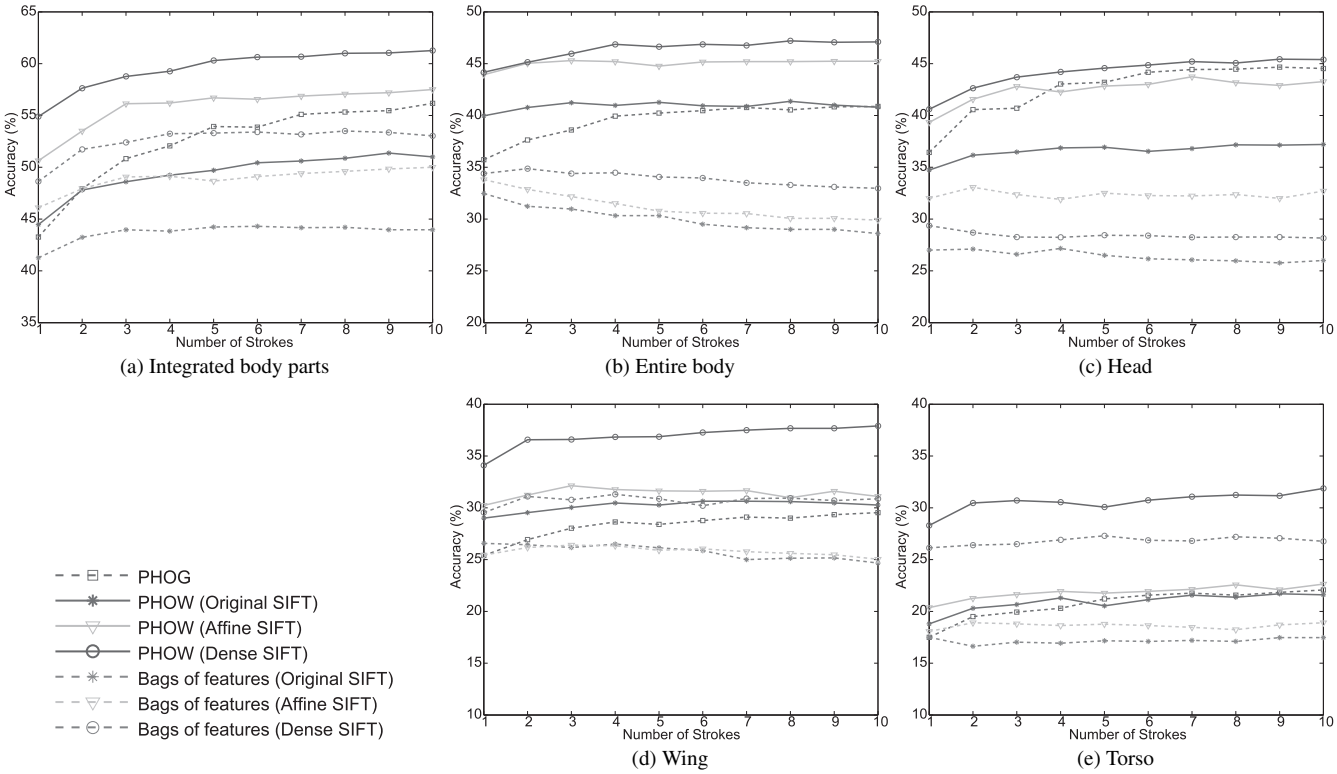
PHOW is similar to PHOG in terms of dividing the target region into multilevel pyramids. Instead of using the orientation of gradients, PHOW calculates the histogram representation of visual words from each cell excluding feature points outside a stroke. Three hundred dimensional vectors are extracted from each cell. This results in a feature vector with dimensionality $300 \sum_{l=0}^{L} 4^l$. The three types of point detectors of original SIFT, dense SIFT, and affine SIFT are also used for PHOW. If $L = 0$, PHOW is equivalent to a bag of features representation. In our experiment, L was set to three for both PHOG and PHOW.

5. Experiments

5.1 Experimental Conditions

The five experimental conditions for classification are summarized below.

- Classifier
  A linear SVM based on one-vs-the rest approach was used for multi-class classification. A software package, the *LIBLINEAR* [25], was used for this purpose.
- Images for training/testing
  Each category of our dataset is composed of 36-112 images. First, 35 images were randomly chosen from each category. The randomly chosen images were separated into 20 training images and 15 test images. Thus the training and the testing samples were not overlapped. Experiments were repeated ten times with different training and testing sets. The mean values and standard deviations for the classification accuracies were evaluated.
- Strokes
  For each training image, multiple manually made and synthesized strokes were used. Two manually made strokes were randomly chosen to synthesize a new synthesized stroke. A blending factor for morphing was randomly determined between 0 and 1. For each testing image, a single manually made stroke was randomly chosen from our datasets.
- Target region
  Two different experiments were performed. (i) **Individual body parts**: The first was where a feature vector was only extracted from one of the four kinds of regions: an entire body, head, wing, and torso. The L2-norm of the feature vector was normalized to 1. (ii) **Integrated body parts**: The second experiment was where all four regions were used in a feature vector. First, four feature vector components were extracted



(a) Integrated body parts (b) Entire body (c) Head (d) Wing (e) Torso

Fig. 6 Relationships between classification accuracies for 20 wild birds and numbers of manually made strokes used for training samples.

from the four kinds of regions, and each of the L1-norms was normalized to one. Then, the four components were concatenated into a single integrated feature vector. The L2-norm of the integrated feature vector was normalized to one.

- Image feature selection
  Seven kinds of image features described in Sect. 4.3 were tested.

5.2 Experimental Results

5.2.1 Results Using Manually Made Strokes

Figure 6 plots the relationships between classification accuracies for the 20 wild birds and numbers of manually made strokes used in the training samples. No synthesized strokes were used for training. The numbers of manually made strokes ranged between 1 to 10. Thus, the total numbers of training samples ranged from 400 up to 4000. The results for *integrated body parts* are given in Fig. 6 (a). In addition, the results for *individual body parts* are presented in Figs. 6 (b)-(d). Seven kinds of image features explained in Sect. 4.3 are compared in these figures.

Table 1 compares the single-stroke results and 10-stroke results. The statistical significances in improvements to classification by expanding the number of training sample were evaluated based on the Wilcoxon signed-rank test at a significance level of 5%. Statistically significant improvements are indicated in bold face.

Table 1 Magnitudes of improvements when numbers of strokes were increased from 1 to 10. Statistically significant improvements are in bold face.

(a) Integrated body parts

Image features	Number of strokes		Improvement
	1	10	
PHOG	43.3 ± 3.7	56.2 ± 2.5	**12.9 ± 2.9**
PHOW (Original SIFT)	44.5 ± 1.9	51.0 ± 2.3	**6.5 ± 2.3**
PHOW (Affine SIFT)	50.6 ± 2.5	57.5 ± 2.0	**6.9 ± 2.3**
PHOW (Dense SIFT)	54.9 ± 2.0	61.3 ± 1.8	**6.4 ± 1.6**
BoF (Original SIFT)	41.3 ± 3.9	44.0 ± 2.2	**2.7 ± 2.8**
BoF (Affine SIFT)	46.1 ± 2.0	50.0 ± 2.6	**3.9 ± 2.6**
BoF (Dense SIFT)	48.6 ± 2.0	53.0 ± 2.3	**4.4 ± 1.6**

BoF: Bags of features

(b) Entire body

Image feature	Number of strokes		Improvement
	1	10	
PHOG	35.7 ± 3.0	40.9 ± 2.7	**5.1 ± 2.1**
PHOW (Original SIFT)	40.0 ± 3.1	40.8 ± 3.0	0.8 ± 1.8
PHOW (Affine SIFT)	43.9 ± 2.2	45.2 ± 1.9	1.3 ± 1.8
PHOW (Dense SIFT)	44.2 ± 2.2	47.1 ± 3.3	**2.9 ± 2.2**
BoF (Original SIFT)	32.5 ± 1.9	28.6 ± 2.8	**−3.9 ± 1.4**
BoF (Affine SIFT)	33.8 ± 0.8	29.9 ± 1.3	**−3.9 ± 1.3**
BoF (Dense SIFT)	34.4 ± 2.6	33.0 ± 2.1	**−1.4 ± 1.6**

5.2.2 Results Using Synthesized Strokes

Figure 7 plots the relationships between accuracies and numbers of augmented strokes obtained with the morphing technique described in Sect. 4.2. In this experiment,

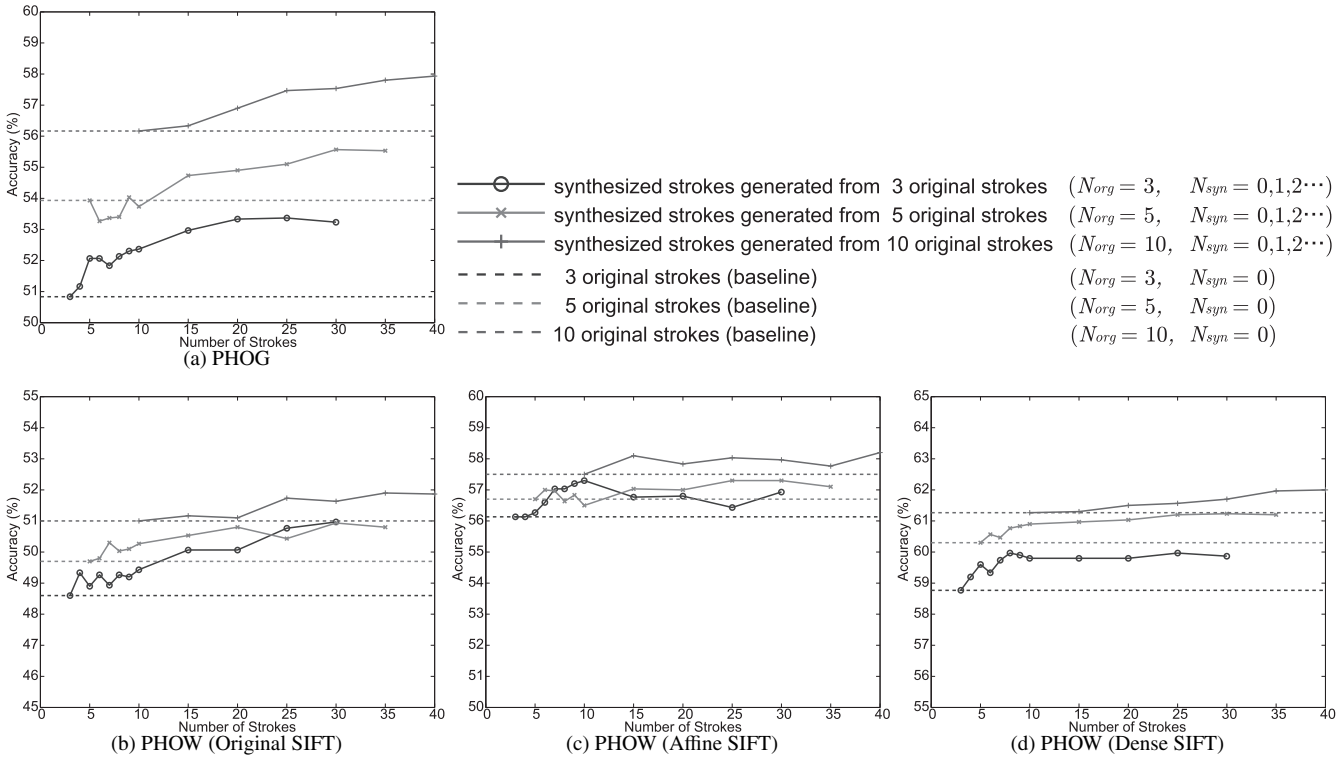


Fig. 7 Relationships between accuracies and numbers of augmented strokes obtained with morphing technique. Results without using synthesized strokes have also been plotted as horizontal lines for comparison.

Table 2 Difference between uses of manually made strokes and augmented strokes, where all four body parts were used. Statistically significant improvements are in bold.

Image feature	Manually made strokes				Augmented strokes				Improvement
	N_{org}	N_{syn}	Sum.	Accuracy	N_{org}	N_{syn}	Sum.	Accuracy	
PHOG	3	0	3	50.8 ± 3.3	3	27	30	53.2 ± 2.4	**2.4 ± 3.2**
	5	0	5	53.9 ± 2.5	5	25	30	55.6 ± 2.5	1.6 ± 3.2
	10	0	10	56.2 ± 2.5	10	20	30	57.5 ± 2.1	1.4 ± 2.2
PHOW (Original SIFT)	3	0	3	48.6 ± 2.7	3	27	30	51.0 ± 2.8	**2.4 ± 1.5**
	5	0	5	49.7 ± 2.2	5	25	30	50.9 ± 2.8	**1.2 ± 1.4**
	10	0	10	51.0 ± 2.3	10	20	30	51.6 ± 2.4	0.6 ± 1.6
PHOW (Affine SIFT)	3	0	3	56.1 ± 1.7	3	27	30	56.9 ± 1.7	0.8 ± 1.5
	5	0	5	56.7 ± 1.8	5	25	30	57.3 ± 2.7	0.6 ± 1.4
	10	0	10	57.5 ± 2.0	10	20	30	58.0 ± 2.1	0.5 ± 0.8
PHOW (Dense SIFT)	3	0	3	58.8 ± 2.4	3	27	30	59.9 ± 2.4	**1.1 ± 1.5**
	5	0	5	60.3 ± 2.3	5	25	30	61.2 ± 2.6	**0.9 ± 1.2**
	10	0	10	61.3 ± 1.8	10	20	30	61.7 ± 2.4	0.4 ± 1.0

N_{syn} synthesized strokes were generated from N_{org} manually made strokes. Three different cases, where N_{org} was set to 3, 5, and 10, were examined. Since the manually made strokes were used together with the synthesized strokes for learning, the number of training samples per image was $N_{org} + N_{syn}$. The results for three cases, where only 3, 5, and 10 manually made strokes were used, have also been plotted as horizontal lines to enable comparison. The results for the four kinds of image features of PHOG and PHOW based on original, affine, and dense SIFT are given in Figs. 7 (a)-(d).

Table 2 summarizes the magnitudes of improvements for PHOG and PHOW in the experiments on *integrated body parts*. As in Table 1, statistically significant improvements are in bold face.

5.3 Discussion

5.3.1 Manually Made Strokes

When image features with spatial information, PHOG, and PHOW were used, classification accuracies were improved as the number of samples increased in most cases as shown in Figs. 6 (a)-(e). The use of the feature vectors of *integrated body parts* yielded particularly significant improvements. PHOG achieved the best improvement of 12.9% on average and three kinds of PHOWs also yielded apparent improvements of 6.4%-6.9% on average. However, bags of features could not obtain as many apparent benefits as

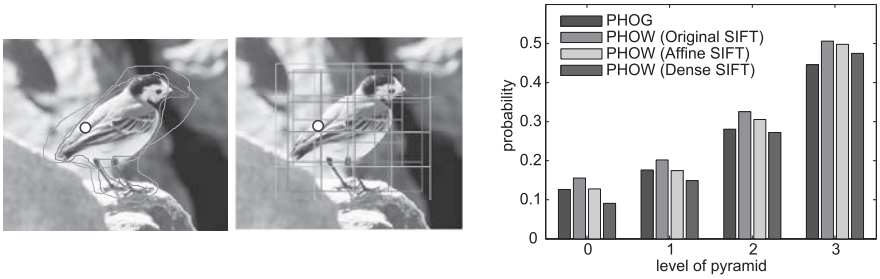


Fig. 8 Quantitative evaluation for variations of feature vectors. (Left) three different strokes and corresponding cells. (Right) probabilities of cases where points of interest were voted into different histogram bins.

PHOG and PHOW. Rather, in many cases where only an *individual body part* was used, accuracies tended to decrease as the number of training samples increased as shown in Figs. 6 (b)-(e).

According to these observations, it can be deduced that incorporating globally spatial information into image features is intrinsically important in the learning of expanded numbers of training samples. Here, a question may arise – why is spatial information so important in improving accuracies? We considered the reasons to be as follows. It is preferable for elements in a feature vector to change globally depending on the deformation of the stroke to obtain higher classification accuracies. Therefore, we evaluated the variations in feature vectors derived by deformation. Figure 8 (left) shows the voting process for creating histogram-based features in PHOG and PHOW. Three different strokes are shown in this figure. For each stroke, the white point is voted into a different histogram bin, because this point belongs to different cells. Figure 8 (right) summarizes the probabilities for where a point of interest (or a pixel in PHOG) was voted into different histogram bins. Interestingly, the probability increased as the pyramid level increased. This means that a higher level for the pyramid globally changes elements of feature vectors. The distributions of feature vectors were expanded in feature space in this way. In other respects, quantization errors in grid partitioning were weakened by ensemble representations of feature vectors in feature space. We believe that this is one reason spatial information is so important.

In an exceptional case, Table 1 (a) lists slight improvements of 2.7%-4.4% on average, even when bags of features were used. However, these features are not true "bags of features" representations, because rough spatial information in the four body parts were incorporated into the integrated feature vector by concatenating the four feature-vector components. This rough spatial information was considered to have had slightly positive effects on classification accuracies.

5.3.2 Synthesized Strokes

Augmenting training samples with synthesized strokes brought about slight improvements in many cases, especially when the synthesized strokes were generated from three manually made strokes. As can been seen from Table 2, no negative effects were found. Therefore, if a limited number of stroke data is available, the use of augmented training samples by synthesizing new strokes is a good alternative to improving classification accuracies.

The degree of improvement depends on the quality of synthesized data. In this work, only two strokes were used to interpolate new contours by using a morphing technique. More sophisticated methods that resemble manually made strokes would improve the accuracy of classification. This is an interesting direction for future work.

The use of a large number of training images will be very important to develop practical applications based on generic object recognition in the future. As a related work, it have been reported that 80 million training samples give impressive improvement on classification accuracy [26]. The way how to collect such a large number of training samples is a significant issue. We could give one of the solutions to the issue.

6. Conclusions

In this paper, we revisited the problem with generic object recognition from the point of view of human-computer interaction. We attempted to improve the classification accuracy by augmenting training samples and obtained the following two conclusions.

1. In most cases, we could improve the classification accuracies by using a large number of the manually made strokes. Image features including spatial information such as PHOG and PHOW tended to have greater apparent positive effects than "bags of features" representations.
2. The use of synthesized segmentation datasets led to a little improvement on the classification accuracies. Although the magnitude of the improvement was a little, this approach gave no negative effect on the classification accuracies even in worst case. If it is difficult to collect a sufficient number of manually made strokes, synthesizing new strokes from them is worth trying. On the other hand, if a large number of the manually made strokes are available, the use of synthesized

strokes will give no improvement.

We believe that it is becoming increasingly important to revisit the problem with generic object recognition from the perspective of human-computer interaction. The best way of separating the recognition task between users and computers should be carefully considered to design practical applications. We believe our work provided new contributions in this direction.

References

[1] P. Viola and M.J. Jones, "Robust real-time face detection," Int. J. Comput. Vis., vol.57, no.2, pp.137–154, May 2004.
[2] H. Harzallah, F. Jurie, and C. Schmid, "Combining efficient object localization and image classification," Proc. ICCV, pp.237–244, 2009.
[3] C.H. Lampert, M.B. Blaschko, and T. Hofmann, "Beyond sliding windows: Object localization by efficient subwindow search," Proc. CVPR, pp.1–8, 2008.
[4] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol.22, no.8, pp.888–905, 2000.
[5] P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik, "From contours to regions: An empirical evaluation," Proc. CVPR, pp.2294–2301, 2009.
[6] C. Gu, J.J. Lim, P. Arbelaez, and J. Malik, "Recognition using regions.," Proc. CVPR, pp.1030–1037, 2009.
[7] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," Proc. CVPR, pp.2169–2178, 2006.
[8] H. Zhang, A.C. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," Proc. CVPR, pp.2126–2136, 2006.
[9] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," Proc. International Conference on Image and Video Retrieval, 2007.
[10] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," Proc. ICCV, pp.1–8, 2007.
[11] L. Fei-Fei, R. Fergus, and P. Perona, "A bayesian approach to unsupervised one-shot learning of object categories," Proc. ICCV, p.1134, 2003.
[12] C. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," Proc. CVPR, pp.951–958, 2009.
[13] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun, "Efficient learning of sparse representations with an energy-based model," Advances in Neural Information Processing Systems 19, pp.1137–1144, 2007.
[14] F. Lauer, C. Suen, and G. Bloch, "A trainable feature extractor for handwritten digit recognition," Pattern Recognit., vol.40, no.6, pp.1816–1824, 2007.
[15] Y. LeCun, and C. Cortes, "THE MNIST DATABASE of handwritten digits." http://yann.lecun.com/exdb/mnist/
[16] J. Zou and G. Nagy, "Visible models for interactive pattern recognition," Pattern Recognit. Lett., vol.28, no.16, pp.2335–2342, 2007.
[17] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification (2nd Edition), 2001.
[18] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results." http://www.pascal-etwork.org/challenges/VOC/voc2009/workshop/index.html.
[19] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.28, pp.594–611, 2006.
[20] "Amazon Mechanical Turk." https://www.mturk.com/mturk/welcome.
[21] C. Scott and R. Nowak, "Robust contour matching via the order-preserving assignment problem," IEEE Trans. Image Process., vol.15, no.7, p.1831, 2006.
[22] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," Proc. Workshop on Statistical Learning in Computer Vision, ECCV, p.22, 2004.
[23] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vis., vol.60, no.2, pp.91–110, 2004.
[24] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," Int. J. Comput. Vis., vol.60, no.1, pp.63–86, 2004.
[25] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "LIBLINEAR: A library for large linear classification," J. Mach. Learn. Res., vol.9, pp.1871–1874, 2008.
[26] A. Torralba, R. Fergus, and W.T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol.30, no.11, pp.1958–1970, 2008.

Mitsuru Ambai received his B.E., M.S., and Ph.D. in information and computer science from Keio University in 2002, 2004 and 2007, respectively. Currently, he is Senior Engineer of research and development group at Denso IT Laboratory, Inc., Tokyo, Japan. His research interests include image processing and computer vision.

Yuichi Yoshida received his B.E., M.S. in engineering science from Osaka University in 2001 and 2003, respectively. From 2003 to 2007, he was a researcher at NTT Corporation. Currently, he is Senior Engineer of research and development group at Denso IT Laboratory, Inc., Tokyo, Japan. His research interests include computer vision and human-computer interaction.